

Received: 23 June 2026; **Revised:** 03 August 2026; **Accepted:** 25 September 2026; **Published:** 30 September 2026

Archs Sci. (2026) Volume 76(3), Pages 46-50, Paper ID 2026306.
<https://doi.org/10.68304/as/76306>

An Extensive Study of Ebola Virus Protein Structure Through Amino Acid Sequences Using Triplet Codes

Thenmozhi Kanaga Ponsindhunesan and S. Arul Mugilan

PG Research Department of Physics, Kamarajar Government Arts College, Surandai-627859, Tamil Nadu, India, Affiliated to Manonmaniam Sundaranar University, Tirunelveli-627012, India
 Corresponding author: Thenmozhi Kanaga Ponsindhunesan (e-mail: thenmozhikanaga@gmail.com).

Abstract Protein-sequence analysis provides an important route for examining the structural organization of viral proteins when sequence composition is considered together with experimentally assigned secondary-structure information. This study investigates Ebola virus proteins by segmenting amino acid sequences into overlapping triplet codes and evaluating the relative enrichment or depletion of those triplets through a normalized deviation parameter. The analysis is designed to identify triplet patterns that occur more or less frequently than expected and to compare these patterns with secondary-structure assignments derived from DSSP. Amino acid triplets are counted using a Python program, observed and expected frequencies are calculated, and the resulting deviation values are normalized to a common 0–1 scale. The same procedure is applied to triplets associated with alpha-helical, beta-sheet, and random-coil regions. The sequence-level analysis identifies several triplets with comparatively high normalized deviations, including methionine-rich and glutamine/tyrosine-containing combinations, while the structure-level analysis indicates that several enriched triplets are associated preferentially with alpha-helical regions. These results provide a descriptive statistical view of triplet usage and its relationship with protein secondary structure. The triplet-based patterns should not be interpreted as direct evidence of biochemical function or therapeutic relevance without independent structural and experimental validation. Nevertheless, the method offers a compact way to compare local amino acid context with structural preference and may complement established sequence- and structure-prediction approaches for Ebola virus proteins.

Index Terms Ebola virus, amino acid triplets, protein sequence analysis, secondary structure, normalized deviation parameter, DSSP

1. Introduction

Ebola virus disease is associated with viruses of the Ebolavirus group and is characterized by severe systemic illness and a historically high case-fatality burden. Structural and sequence-based studies of Ebola virus proteins have therefore received sustained attention because viral proteins mediate entry, replication, assembly, immune evasion, and interactions with host factors. Cong, Pei, and Grishin performed predictive and comparative analyses of Ebolavirus proteins, integrating sequence variation with available three-dimensional structures and identifying putative functional regions in incompletely characterized viral proteins [1]. Such structural information provides a basis for interpreting sequence differences in relation to viral mechanisms and for identifying regions that may warrant further biochemical investigation.

The clinical importance of Ebola virus protein research is also evident from the development of therapeutic interventions. The randomized controlled trial reported by Mulangu et al. demonstrated that specific antibody-based treatments can substantially improve outcomes in Ebola virus disease [2]. Although that clinical study does not establish protein structural

motifs directly, it illustrates why rigorous characterization of viral proteins and their functional regions remains relevant to therapeutic research.

Protein secondary-structure prediction has a long methodological history. Chou and Fasman developed early statistical rules for estimating secondary-structure propensity directly from amino acid sequences [3], while statistical analysis of protein-sequence patterns has also been used to characterize biologically meaningful residue distributions [4]. Earlier conformational-parameter work by Chou and Fasman quantified amino acid preferences for alpha-helical, beta-sheet, and random-coil regions [5]. These methods established the principle that local sequence composition contains information related to protein conformation.

The GOR approach formulated secondary-structure prediction using information-theoretic reasoning and local sequence context [6]. Rost, Sander, and Schneider subsequently developed the PHD system, which integrated sequence profiles and neural-network prediction [7]. Position-specific scoring matrices further improved secondary-structure prediction by incorporating evolutionary information, as demonstrated by Jones [8]; PSI-BLAST provided an efficient framework for

generating such sequence profiles from iterative database searches [9].

Machine-learning methods extended these approaches by learning nonlinear relationships between sequence context and secondary structure [10]. Neural-network methods include the work of Holley and Karplus [11], Qian and Sejnowski [12], Kneller, Cohen, and Langridge [13], Malekpour et al. [14], and Qu et al. [15]. Hidden Markov and hidden semi-Markov approaches were developed by Asai et al. [16], Won et al. [17], and Aydin, Altunbasak, and Borodovsky [18]. Support-vector-machine methods were reported by Kim and Park [19], Ward et al. [20], Guo et al. [21], and Hua and Sun [22]. Fuzzy K-nearest-neighbor classification has also been applied to protein secondary-structure class prediction [23]. Collectively, these studies show that sequence-based structural prediction can be approached using statistical, evolutionary, probabilistic, and machine-learning frameworks.

The present work takes a different but complementary descriptive approach. Rather than training a predictive classifier, it partitions Ebola virus protein sequences into overlapping amino acid triplets and quantifies enrichment or depletion relative to an expected frequency. The resulting deviation parameter is normalized to allow comparison across triplets. The same principle is then applied to DSSP-derived secondary-structure categories to examine whether particular triplets are preferentially associated with alpha-helices, beta-sheets, or random structures.

The main objective is therefore to determine whether specific amino acid triplets are statistically over-represented in the analyzed Ebola virus protein sequences and whether these same triplets show distinctive secondary-structure preferences. The study does not attempt to infer biological function solely from sequence frequency. Functional interpretations in the Discussion are presented as hypotheses requiring independent structural, biochemical, or mutational validation.

II. Materials and Methods

A. Sequence Analysis

Ebola virus protein structures and associated sequences were obtained from the Protein Data Bank resource used in the study [24]. Protein sequences were processed using a Python program that enumerates overlapping amino acid triplets. For a sequence of length L , the sliding triplet procedure produces $L - 2$ overlapping triplets, with consecutive windows shifted by one residue.

For each triplet, the number of occurrences in the analyzed Ebola virus sequence set was counted. The observed frequency was then calculated relative to the total number of triplets. An expected frequency was calculated using the background protein-sequence frequencies defined in the study. The observed and expected frequencies were used to derive a percentage deviation for each triplet.

This procedure provides a descriptive measure of triplet enrichment. A positive deviation indicates that a triplet occurs more frequently than expected, while a negative deviation indicates under-representation. Because raw deviations can

span different numerical ranges, min-max normalization was used to place the values on a common scale between 0 and 1.

B. Structure Analysis

Secondary-structure assignments were obtained from DSSP, the established dictionary-of-secondary-structure procedure based on hydrogen-bonding and geometric criteria [25]. The analyzed residues were grouped into the structural categories used in this study: alpha-helix, beta-sheet, and random structure. A Python program was used to count amino acid residues and triplet occurrences associated with each structural category.

Observed and expected frequencies were calculated for the structural analysis using the same general statistical procedure as in the sequence analysis. Deviations were computed and normalized, allowing the structural preferences of selected triplets to be compared across alpha-helical, beta-sheet, and random regions.

Figure 1 summarizes the complete workflow, including sequence acquisition, triplet counting, expected-frequency calculation, deviation analysis, normalization, DSSP-based structural classification, and interpretation.

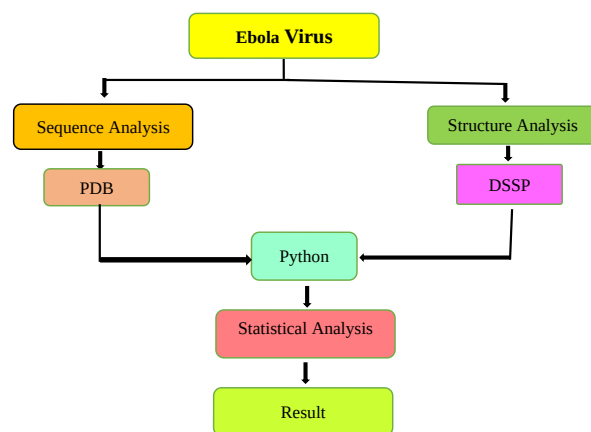


Figure 1: Workflow for Amino Acid Triplet and Secondary-Structure Analysis

C. Statistical Analysis

For the sequence analysis, the observed triplet frequency is defined as

$$f_{\text{obs}}(t) = \frac{n_{\text{Ebola}}(t)}{N_{\text{Ebola}}}, \quad (1)$$

where $n_{\text{Ebola}}(t)$ is the number of occurrences of triplet t in the analyzed Ebola virus protein sequences and N_{Ebola} is the total number of overlapping triplets in those sequences.

The expected frequency is defined as

$$f_{\text{exp}}(t) = \frac{n_{\text{background}}(t)}{N_{\text{background}}}, \quad (2)$$

where $n_{\text{background}}(t)$ and $N_{\text{background}}$ refer to the corresponding triplet count and total triplet count in the background protein set used by the study.

The percentage deviation parameter is

$$D(t) = \frac{f_{\text{obs}}(t) - f_{\text{exp}}(t)}{f_{\text{exp}}(t)} \times 100. \quad (3)$$

A positive value indicates over-representation relative to the background expectation, whereas a negative value indicates under-representation. A value of zero indicates agreement between the observed and expected frequencies. The use of residue-frequency and triplet-deviation analysis is consistent with earlier statistical studies of amino acid composition and higher-order sequence patterns [26]–[28].

For cross-triplet comparison, the deviation parameter is normalized by min–max scaling:

$$D_{\text{norm}}(t) = \frac{D(t) - D_{\min}}{D_{\max} - D_{\min}}. \quad (4)$$

Here, $D_{\min}$ and $D_{\max}$ are the minimum and maximum deviation values in the comparison set. The normalized value therefore lies between 0 and 1. A value near 1 identifies a triplet with a deviation close to the maximum observed in the dataset, while a value near 0 identifies a triplet close to the minimum. The same normalization procedure is applied to the secondary-structure analysis [26]–[28].

It is important to note that min–max normalization ranks triplets relative to the analyzed dataset; it does not by itself provide a probability, significance level, or effect-size estimate. The normalized values are therefore interpreted descriptively.

III. Results and Discussion

A. Amino Acid Triplet Distribution

Figure 2 presents the normalized deviation values for the selected amino acid triplets identified in the Ebola virus sequence analysis.

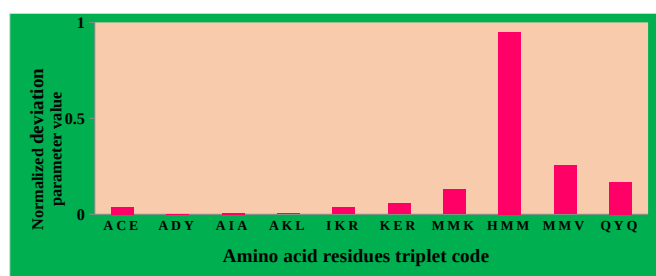


Figure 2: Normalized Deviation Analysis of Amino Acid Triplet Codes in Ebola Virus Proteins

Most analyzed triplets have normalized deviation values close to the lower end of the scale, whereas a smaller subset displays markedly larger values. In the plotted results, HMM, MMV, MMM, and QYQ are among the triplets with comparatively strong deviations. These values indicate enrichment relative to the study's background expectation; they do not, by themselves, establish conservation, essentiality, or biological function.

The biochemical composition of the enriched triplets nevertheless provides hypotheses that can be tested structurally. ACE contains alanine, cysteine, and glutamate. Cysteine can participate in disulfide bonds when positioned appropriately in an oxidizing structural environment, while glutamate can contribute to electrostatic interactions. ADY contains aspartate and tyrosine, residues that can participate in hydrogen bonding or regulatory chemistry depending on their structural context. AIA is predominantly hydrophobic and could be compatible with buried or membrane-associated environments, although sequence enrichment alone cannot determine the corresponding structural location.

AKL combines alanine, lysine, and leucine. Lysine can participate in electrostatic interactions, whereas leucine has a strong general compatibility with hydrophobic and helical environments. IKR and KER contain multiple charged residues and may therefore occur in solvent-exposed or interaction-prone regions, but the observed triplet frequency cannot identify a specific nucleic-acid-binding or localization function without independent evidence.

Methionine-rich triplets deserve careful interpretation. MMK, HMM, MMM, and MMV contain two or three methionine residues and show comparatively high deviation values in the plotted analysis. Methionine is the initiating residue in canonical protein translation, but an internal methionine-rich triplet does not imply that the corresponding region is itself a translation-initiation site. The enrichment may instead reflect local compositional characteristics of the analyzed proteins. Similarly, histidine in HMM can participate in catalytic or metal-binding environments in some proteins, but no enzymatic function can be assigned to HMM from frequency analysis alone.

QYQ combines glutamine and tyrosine. Both residues can participate in hydrogen-bonding networks, while tyrosine may undergo phosphorylation in suitable cellular contexts. These biochemical properties make QYQ structurally interesting, but experimental validation would be required before assigning a regulatory role.

Earlier triplet-frequency studies provide the statistical context for interpreting such enrichments [27], [28]. The principal result of Figure 2 is therefore the identification of non-uniform triplet usage in the analyzed Ebola virus sequences. Functional interpretations should remain secondary to that quantitative observation.

B. Secondary-Structure Preferences

Figure 3 presents the normalized deviations of selected triplets across alpha-helix, beta-sheet, and random-structure categories.

The results indicate that several triplets exhibit relatively strong alpha-helical preferences, whereas others are associated more strongly with beta-sheet or random-structure categories. Such observations are consistent with the broader principle that amino acid propensities vary across secondary-structure environments. Position-dependent alpha-helical propensities have been demonstrated experimentally

and statistically [29], and amino acid preferences for helices and beta-sheets also depend on the overall protein fold and structural context [30]. The central role of alpha-helices and beta-sheets as principal protein structural features is well established [31].

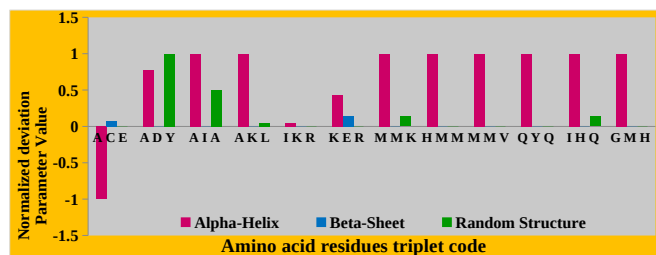


Figure 3: Secondary-Structure Preference Analysis of Ebola Virus Amino Acid Triplet Codes

In the present analysis, AIA and AKL show relatively strong alpha-helical tendencies. This pattern is qualitatively compatible with the known helix-forming properties of alanine and leucine, although the observed triplet-level deviation should not be equated directly with an experimentally measured helix propensity. KER shows moderate representation across more than one structural class, suggesting that charged-residue context may be compatible with multiple local geometries.

MMK and other methionine-rich triplets also show pronounced alpha-helical deviations in the plotted data. HMM and MMM are among the strongest examples in the selected set. This may indicate that methionine-rich local contexts occur preferentially within helical regions of the analyzed Ebola virus structures. The result is descriptive and should be confirmed by mapping individual triplet occurrences back to specific PDB residues and protein domains.

Triplets such as ADY, IQH, and GMH display comparatively stronger random-structure components in the analysis. Random or coil assignments can correspond to loops, linkers, turns, flexible termini, or disordered segments, but the DSSP-derived category used here does not provide a direct measure of intrinsic disorder or biological flexibility. Accordingly, the structural interpretation should remain tied to the secondary-structure classification actually analyzed.

The simultaneous use of sequence enrichment and DSSP-based structural preference is the main methodological contribution of the study. A triplet that is enriched at the sequence level and also concentrated within a specific structural category becomes a candidate for more detailed examination. Such follow-up could include residue-level mapping, solvent accessibility, evolutionary conservation, molecular dynamics, mutational analysis, or comparison across Ebolavirus species.

C. Methodological Interpretation and Limitations

The normalized deviation parameter provides an intuitive ranking of triplet enrichment, but several limitations should be acknowledged. First, overlapping triplets are not statistically

independent because adjacent windows share two residues. Conventional significance tests that assume independent observations would therefore require modification before being applied directly to these counts.

Second, the expected-frequency model strongly influences the deviation. If the background protein set differs substantially in amino acid composition from Ebola virus proteins, enrichment may partly reflect overall composition rather than triplet-specific organization. Future work should compare multiple background models, including independent-residue expectations, composition-matched shuffled sequences, and related viral proteins.

Third, min-max normalization depends on the extreme values in the dataset. Adding or removing a strongly enriched triplet can change the normalized values of all other triplets even if their raw frequencies remain unchanged. Reporting both raw deviation and normalized deviation would therefore improve reproducibility.

Fourth, the present structural analysis groups DSSP assignments into three broad categories. A more detailed analysis could retain individual DSSP states and distinguish alpha-helices, 3_{10} helices, beta-strands, turns, bends, and unstructured states. This would provide a finer structural interpretation.

Finally, enrichment does not establish function. Claims concerning membrane association, phosphorylation, catalysis, nucleic-acid binding, viral replication, or therapeutic targeting require direct supporting evidence. The current method is most appropriately viewed as a screening strategy that identifies sequence and structure patterns for subsequent validation.

IV. Conclusion

This study applies an overlapping amino acid triplet analysis to Ebola virus protein sequences and combines the resulting frequency deviations with DSSP-based secondary-structure assignments. The approach identifies non-uniform triplet usage and reveals that several selected triplets, particularly methionine-rich combinations, show comparatively strong enrichment and alpha-helical preference in the analyzed dataset.

The normalized deviation parameter provides a simple standardized measure for comparing triplets whose raw deviations differ in scale. When applied separately to sequence occurrence and secondary-structure categories, it allows candidate triplets to be prioritized for more detailed structural analysis. The results are therefore useful as a descriptive screening layer rather than as direct proof of protein function.

Future work should map enriched triplets to individual Ebola virus proteins and PDB structures, compare results across Ebolavirus species, use composition-matched sequence controls, and evaluate statistical uncertainty through permutation or resampling methods. Integrating evolutionary conservation, solvent accessibility, structural-domain annotation, and experimentally characterized functional sites would further clarify whether the identified triplet patterns have biological significance.

Author Contributions

T.K.P. and S.A.M. contributed equally to the conceptualization, methodology, software implementation, sequence analysis, structural analysis, interpretation, visualization, manuscript preparation, and review and editing. Both authors have read and approved the final manuscript.

Funding

No specific external funding information was provided for this study.

Data Availability

The structural and sequence information analyzed in this study was obtained from publicly accessible PDB and DSSP resources cited in the manuscript. The derived triplet counts, processed data, and analysis scripts are available from the corresponding author upon reasonable request.

Ethics Statement

Ethical approval and informed consent were not required because this study used publicly available protein-sequence and structural data and did not involve human participants, identifiable personal data, or animal experimentation.

Conflict of Interest

The authors declare no conflicts of interest.

References

- [1] Cong, Q., Pei, J., & Grishin, N. V. (2015). Predictive and comparative analysis of Ebolavirus proteins. *Cell Cycle*, 14(17), 2785–2797.
- [2] Mulangu, S., Dodd, L. E., Davey, R. T., Jr., Mbaya, O. T., Proschan, M., Mukadi, D., Manzo, M. L., Nzolo, D., Oloma, A. T., Ibanda, A., Ali, R., Coulibaly, S., Levine, A. C., Grais, R., Diaz, J., Lane, H. C., & Muyembe-Tamfum, J.-J. (2019). A randomized, controlled trial of Ebola virus disease therapeutics. *The New England Journal of Medicine*, 381(24), 2293–2303.
- [3] Chou, P. Y., & Fasman, G. D. (1978). Prediction of the secondary structure of proteins from their amino acid sequence. *Advances in Enzymology and Related Areas of Molecular Biology*, 47, 45–148.
- [4] Brendel, V., Bucher, P., Nourbakhsh, I. R., Blaisdell, B. E., & Karlin, S. (1992). Methods and algorithms for statistical analysis of protein sequences. *Proceedings of the National Academy of Sciences of the United States of America*, 89(6), 2002–2006.
- [5] Chou, P. Y., & Fasman, G. D. (1974). Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. *Biochemistry*, 13(2), 211–222.
- [6] Garnier, J., Gibrat, J.-F., & Robson, B. (1996). GOR method for predicting protein secondary structure from amino acid sequence. *Methods in Enzymology*, 266, 540–553.
- [7] Rost, B., Sander, C., & Schneider, R. (1994). PHD—An automatic mail server for protein secondary structure prediction. *Computer Applications in the Biosciences*, 10(1), 53–60.
- [8] Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology*, 292(2), 195–202.
- [9] Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, 25(17), 3389–3402.
- [10] Yoo, P. D., Zhou, B. B., & Zomaya, A. Y. (2008). Machine learning techniques for protein secondary structure prediction: An overview and evaluation. *Current Bioinformatics*, 3(2), 74–86.
- [11] Holley, L. H., & Karplus, M. (1989). Protein secondary structure prediction with a neural network. *Proceedings of the National Academy of Sciences of the United States of America*, 86(1), 152–156.
- [12] Qian, N., & Sejnowski, T. J. (1988). Predicting the secondary structure of globular proteins using neural network models. *Journal of Molecular Biology*, 202(4), 865–884.
- [13] Kneller, D. G., Cohen, F. E., & Langridge, R. (1990). Improvements in protein secondary structure prediction by an enhanced neural network. *Journal of Molecular Biology*, 214(1), 171–182.
- [14] Malekpour, S. A., Naghizadeh, S., Pezeshk, H., Sadeghi, M., & Eslahchi, C. (2009). Protein secondary structure prediction using three neural networks and a segmental semi-Markov model. *Mathematical Biosciences*, 217(2), 145–150.
- [15] Qu, W., Sui, H., Yang, B., & Qian, W. (2011). Improving protein secondary structure prediction using a multi-modal BP method. *Computers in Biology and Medicine*, 41(10), 946–959.
- [16] Asai, K., Hayamizu, S., & Handa, K. (1993). Prediction of protein secondary structure by the hidden Markov model. *Computer Applications in the Biosciences*, 9(2), 141–146.
- [17] Won, K.-J., Hamelryck, T., Prügler-Bennett, A., & Krogh, A. (2007). An evolutionary method for learning HMM structure: Prediction of protein secondary structure. *BMC Bioinformatics*, 8, 357.
- [18] Aydin, Z., Altunbasak, Y., & Borodovsky, M. (2006). Protein secondary structure prediction for a single-sequence using hidden semi-Markov models. *BMC Bioinformatics*, 7, 178.
- [19] Kim, H., & Park, H. (2003). Protein secondary structure prediction based on an improved support vector machines approach. *Protein Engineering*, 16(8), 553–560.
- [20] Ward, J. J., McGuffin, L. J., Buxton, B. F., & Jones, D. T. (2003). Secondary structure prediction with support vector machines. *Bioinformatics*, 19(13), 1650–1655.
- [21] Guo, J., Chen, H., Sun, Z., & Lin, Y. (2004). A novel method for protein secondary structure prediction using dual-layer SVM and profiles. *Proteins: Structure, Function, and Bioinformatics*, 54(4), 738–743.
- [22] Hua, S., & Sun, Z. (2001). A novel method of protein secondary structure prediction with high segment overlap measure: Support vector machine approach. *Journal of Molecular Biology*, 308(2), 397–407.
- [23] Tan, Y. T., & Rosdi, B. A. (2015). FPGA-based hardware accelerator for the prediction of protein secondary class via fuzzy K-nearest neighbors with Lempel–Ziv complexity based distance measure. *Neurocomputing*, 148, 409–419.
- [24] Joosten, R. P., te Beek, T. A. H., Krieger, E., Hekkelman, M. L., Hooft, R. W. W., Schneider, R., Sander, C., & Vriend, G. (2011). A series of PDB related databases for everyday needs. *Nucleic Acids Research*, 39(Database issue), D411–D419.
- [25] Kabsch, W., & Sander, C. (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12), 2577–2637.
- [26] Rani, M., & Mitra, C. K. (1995). Correlation analysis of frequency distributions of residues in proteins. *Journal of Biosciences*, 20(1), 7–16.
- [27] Veluraja, K., & Mugilan, S. A. (1997). Amino acid doublets and triplets in protein sequences: A database analysis. *Current Science*, 72(8), 572–577.
- [28] Mugilan, S. A., & Veluraja, K. (2000). Generation of deviation parameters for amino acid singlets, doublets and triplets from three-dimensional structures of proteins and its implications for secondary structure prediction from amino acid sequences. *Journal of Biosciences*, 25(1), 81–91.
- [29] Engel, D. E., & DeGrado, W. F. (2004). Amino acid propensities are position-dependent throughout the length of α -helices. *Journal of Molecular Biology*, 337(5), 1195–1205.
- [30] Fujiwara, K., Toda, H., & Ikeguchi, M. (2012). Dependence of α -helical and β -sheet amino acid propensities on the overall protein fold type. *BMC Structural Biology*, 12, 18.
- [31] Eisenberg, D. (2003). The discovery of the α -helix and β -sheet, the principal structural features of proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 100(20), 11207–11210.

...